Pitfalls (& Solutions) for Assigning Molecular Function from Structural Similarities:

Mechanistically Diverse Enzyme Superfamilies
Definitions of function: E.C. & GO

E.C. (Enzyme Nomenclature) system (http://www.chem.qmul.ac.uk/iubmb/enzyme/)
> 4-digit hierarchical system for naming chemical reactions
> not correlated with structure-based hierarchies

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>-.</td>
<td>--</td>
</tr>
<tr>
<td>2.</td>
<td>-.</td>
<td>--</td>
</tr>
<tr>
<td>3.</td>
<td>-.</td>
<td>--</td>
</tr>
<tr>
<td>4.</td>
<td>-.</td>
<td>--</td>
</tr>
<tr>
<td>5.</td>
<td>-.</td>
<td>--</td>
</tr>
<tr>
<td>5. 1.</td>
<td>-.</td>
<td>--</td>
</tr>
<tr>
<td>5. 1. 2.</td>
<td>-.</td>
<td>--</td>
</tr>
<tr>
<td>6.</td>
<td>-.</td>
<td>--</td>
</tr>
</tbody>
</table>

- OXIDOREDUCTASES
- TRANSFERASES
- HYDROLASES
- LYASES
- ISOMERASES
- RACEMASES AND EPIMERASES
  - ACTING ON HYDROXY ACIDS/DERIVATIVES
    - MANDELA RACEMASE
- LIGASES

Gene Ontology (GO) system (http://www.geneontology.org/)
> 3 structured, controlled vocabularies describe gene products in terms of their associated biological processes, cellular components and molecular functions
  - for this study, tracks well with E.C. data
we can see further using structure than we can with sequence

(Chothia & Lesk (1986) Embo J 5: 823-6)
> but homologs that can only be seen/verified at the structural level
have also frequently diverged to mediate very different overall
functions

> how to discriminate and accurately annotate function from
structural information?
HAD superfamily includes many different overall reactions, including ATPases, β-phosphoglucomutase, phosphatases, phosphonatases, haloacid dehalogenase.

- All members of the SF use the same active site machinery to perform a similar fundamental function, hydrolysis & phosphoryl group transfer via hydrolytic nucleophilic substitution.
- We call these types of superfamilies “functionally [or mechanistically] diverse”.
- Annotation transfer from structural similarities in these types of SFs is problematic because they look similar but perform quite different rxns.
Annotation transfer in functionally diverse enzyme superfamilies

› Requires models for the relationship between evolutionary divergence in structure and changes in function to tell us what characteristics of function can be associated with conserved elements of sequence/structure

› Annotation transfer of function from sequence/structural similarity should use descriptions of function that can be explicitly associated with conserved sequence & structural features of homologous proteins

› Sorting out superfamily/family relationships is not trivial for these types of enzyme SFs
Retro (substrate-constrained) evolution

The original model for evolution of enzyme function predicts that conserved elements of structure are associated with the ability to bind the same ligand

Horowitz, PNAS 31 (1945) 153-157;
Horowitz, in Evolving Genes & Proteins (1965) 15-23

- structural elements involved in substrate binding are conserved across divergent proteins while new chemistries evolve
  - typically represented by pairs of structures/functions rather than large superfamilies of many different structures/functions
Chemistry-constrained evolution: Mechanistically diverse superfamilies

Structural elements involved in mediating a fundamental aspect of chemistry are conserved across divergent proteins while the ability to bind new substrates & perform different overall reactions evolves.

Babbitt & Gerlt, JBC 272: 30591-30594 (1997)
Active site constrained evolution: Functionally distinct suprafamilies

Active site structure is dominant: Active site structure is conserved but used for different mechanistic steps in the overall catalytic mechanisms, e.g., no common partial reaction


- OMP decarboxylase - no metal, mechanism avoids formation of an unstable anion intermediate

- KGP decarboxylase - metal assisted stabilization of an enediolate anion intermediate

More suprafamilies are being discovered as our ability to identify more distant structural relationships improves
Mechanistically Diverse SFs: Enolase Superfamily

Structural elements involved in mediating chemistry are conserved across divergent proteins while the ability to bind new substrates evolves.

\[ \text{Mandelate Racemase} \quad \text{Muconate Lactonizing Enzyme} \]
> similarities in the active sites map to a conserved partial reaction rather than to the overall reactions
Using the principles of chemistry-constrained evolution to improve function prediction from structure

because all of these proteins show conserved structures/active sites, it is easy to make the mistaken inference that they perform the same overall function

> 1000+ sequences ranging from <15-65% identity between families that perform different overall functions

> ~50% of these sequences are still of unknown function
the conserved elements of all of these active sites perform the same function: abstraction of a proton alpha to a carboxylic acid

Babbitt et al., Biochem., 35 16489-16501 (1996)
across the SF, the non-conserved elements of these active sites specify the differences in the functions that these enzymes perform
What is the function of Orf587?

S-substrate

\[ \text{Orf587} \quad \text{QIGFDTFKLNGCEEL} \]
\[ \text{MLE} \quad \text{IRRRHVFKLKITGADP} \]
\[ \text{MR} \quad \text{ELGFRAVKTKIGYP} \]

Metal Binding

\[ \text{Orf587} \quad \text{LDHFGRVSAPMAKVLIKEPEPYRPLFIEVPLAEQ-AEYYPKLAQTH---IPLAAGER} \]
\[ \text{MLE} \quad \text{VDVNQYWDSEQAIRACQVLGDNGIDLIEQPISRIN-RGGQVRNLQRSP---APIMADESI} \]
\[ \text{MR} \quad \text{VDYNQSLDVPAIKRSQALQEGVTWIEEPTLQHD-YEGHRQIQSKLN---VPVQME} \]

R-substrate

\[ \text{Orf587} \quad \text{QPD-----SHAGGITECYKIAGMAEAYDVTLAPHCP--LGPIALAACLHFVSYNAVLEQS} \]
\[ \text{MLE} \quad \text{ALKI-----AKNGGPRAVLRTANIAEAAGIGLYGTMLEGAIOTLASAHFLTLRQLTGWTEL} \]
\[ \text{MR} \quad \text{MPDA-----MKIGGVTGWIRASALAQQFGIPMSH-------LFQEISAHLAATP-TAHWHEL} \]
Superfamily model: proton abstraction
Subgroup clustering: stereospecificity of rxn
Metabolic context: substrate specificity

Several large superfamilies that fit the chemistry constrained model have now been characterized in detail.

- **Enolase**: metal dependent abstraction of α-protons of carboxylic acids
- **Vicinal oxygen chelate**: stabilization of diverse oxyanion intermediates
- **Crotonase**: stabilization of oxyanion intermediates derived from thioesters
- **Haloacid dehalogenase**: hydrolytic nucleophilic substitution
How many functionally diverse enzyme superfamilies in the protein universe?

> 49% of superfamilies (with >= 2 families per superfamily) do not have any conserved EC annotations
> 18% of all SCOPEC superfamilies do not have any conserved EC annotations
> analogous results with GO assignments
> counting SFs by whether all members can be identified using a single 3D motif representing highly conserved active site residues gives a much lower number.

Because E.C. describes only overall functions, it is problematic for mapping conserved function to the similarities in SF members' structures/active sites.

<table>
<thead>
<tr>
<th>Superfamily</th>
<th>Fundamental partial reaction/chemical capability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Enolase</td>
<td>Metal dependent abstraction of ( \alpha )-protons of carboxylic acids to form stabilized enolate intermediates</td>
</tr>
<tr>
<td>EC number</td>
<td>Overall reaction</td>
</tr>
<tr>
<td>4.2.1.6</td>
<td>galactonate dehydratase</td>
</tr>
<tr>
<td>4.2.1.11</td>
<td>enolase</td>
</tr>
<tr>
<td>4.2.1.40</td>
<td>glucarate dehydratase</td>
</tr>
<tr>
<td>4.2.1.-</td>
<td>( \alpha )-succinylbenzoate-CoA synthase</td>
</tr>
<tr>
<td>4.3.1.2</td>
<td>methylaspartate ammonia-lyase</td>
</tr>
<tr>
<td>5.1.2.3</td>
<td>mandelate racemase</td>
</tr>
<tr>
<td>5.5.1.1</td>
<td>muconate lactonizing enzyme</td>
</tr>
<tr>
<td>Crotonase</td>
<td>Stabilization of oxyanion intermediates derived from thioesters</td>
</tr>
<tr>
<td>EC number</td>
<td>Overall reaction</td>
</tr>
<tr>
<td>3.1.2.4</td>
<td>3-hydroxyisobutyryl-CoA hydrolase</td>
</tr>
<tr>
<td>3.4.21.92</td>
<td>ATP-dependent Clp protease</td>
</tr>
<tr>
<td>3.8.1.6</td>
<td>4-chlorobenzoyl-CoA dehalogenase</td>
</tr>
<tr>
<td>4.1.1.41</td>
<td>methymalonyl-CoA decarboxylase</td>
</tr>
<tr>
<td>4.1.3.36</td>
<td>naphthoate synthase</td>
</tr>
<tr>
<td>4.2.1.17</td>
<td>enoyl-CoA hydratase (crotonase)</td>
</tr>
<tr>
<td>5.3.3.-</td>
<td>( \Lambda^{1,5,6} \Lambda^{2,4} )-diencyclohexyl-CoA isomerase</td>
</tr>
<tr>
<td>Haloacid dehalogenase</td>
<td>Hydrolysis, phosphoryl group transfer via hydrolytic nucleophilic substitution</td>
</tr>
<tr>
<td>3.1.3.3</td>
<td>phosphoserine phosphatase</td>
</tr>
<tr>
<td>3.1.3.15</td>
<td>histidinol phosphatase</td>
</tr>
<tr>
<td>3.11.1.1</td>
<td>phosphonatase</td>
</tr>
<tr>
<td>3.1.3.18</td>
<td>phosphoglycolate phosphatase</td>
</tr>
<tr>
<td>3.8.1.2</td>
<td>haloacid dehalogenase</td>
</tr>
<tr>
<td>5.4.2.6</td>
<td>( \beta )-phosphoglucomutase</td>
</tr>
</tbody>
</table>

> for these systems, annotation transfer by sequence or structural similarity to a protein of characterized E.C. class is suspect

> for example, 25% of sequences annotated as muconate lactonizing sequences in Genbank are likely misannotated

- some are also misannotated in SwissProt & derivative db’s

<table>
<thead>
<tr>
<th>gi number</th>
<th>Evidence code</th>
<th>HMM</th>
<th>Cat operon</th>
</tr>
</thead>
<tbody>
<tr>
<td>1633162</td>
<td>STRC</td>
<td>1e-273</td>
<td>3</td>
</tr>
<tr>
<td>5915882</td>
<td>IDA</td>
<td>4e-238</td>
<td>3</td>
</tr>
<tr>
<td>151123</td>
<td>ISO</td>
<td>1e-267</td>
<td>3</td>
</tr>
<tr>
<td>23491535</td>
<td>ISO</td>
<td>1e-263</td>
<td>3</td>
</tr>
<tr>
<td>4579699</td>
<td>ISO</td>
<td>3e-274</td>
<td>3</td>
</tr>
<tr>
<td>7437422</td>
<td>ISO</td>
<td>1e-272</td>
<td>3</td>
</tr>
<tr>
<td>5915881</td>
<td>ISO</td>
<td>1e-262</td>
<td>3</td>
</tr>
<tr>
<td>9948563</td>
<td>IEA</td>
<td>5e-263</td>
<td>3</td>
</tr>
<tr>
<td>13476122</td>
<td>IEA</td>
<td>5e-17</td>
<td>0</td>
</tr>
<tr>
<td>23100420</td>
<td>IEA</td>
<td>4e-34</td>
<td>0</td>
</tr>
<tr>
<td>15615568</td>
<td>IEA</td>
<td>1e-20</td>
<td>0</td>
</tr>
<tr>
<td>15642781</td>
<td>IEA</td>
<td>1e-16</td>
<td>0</td>
</tr>
<tr>
<td>17231024</td>
<td>IEA</td>
<td>1e-13</td>
<td>0</td>
</tr>
<tr>
<td>29346723</td>
<td>IEA</td>
<td>1e-14</td>
<td>0</td>
</tr>
<tr>
<td>23100298</td>
<td>IEA</td>
<td>1e-17</td>
<td>0</td>
</tr>
</tbody>
</table>
Sequence, structure, experimental validation, including mechanistic information used to curate the structure-function linkage for each SF.

http://sfld.rbvi.ucsf.edu/

Different SF sequences: low % identity

Subset of active site residues conserved in all members of a SF perform a conserved partial rxn

Conserved subset of active site residues correlate with subgroup-specific differences in function

Family level: Family-specific subsets of residues mediate specificity & overall rxns

% ID within families >% ID between families

Pegg, SC., et al. Pac Symp Biocomput, 2005
Define function not only in terms of the overall reaction but also in terms of the individual partial reactions (or chemical capabilities) that make up each overall reaction.

> a “Gold Standard” set of mechanistically diverse SFs will soon be available for download from the SFLD
  - test set for methods development in linking sequences/structures to SF/subgroup/family specific definitions of function
SFLD analysis tools

> SF, subgroup, & family-specific HMMs for classifying new sequences & re-classifying sequences misannotated in the databases

> highly curated multiple alignments/automated alignments & dendograms

> sequences & structures clustered into subgroups/families within each SF provide context for interpreting alignments of new sequences to a family, subgroup, or SF
  - multiple alignments, dendograms
  - evidence codes, Pubmed abstracts, critical active site residues required of a cluster at each level of the hierarchy
  - modeled structures (courtesy of Sali lab’s Modbase)

> active site templates for finding/classifying new structures of unknown function (in process)

> structure-based comparison (in process)
In the pipeline:

- 5+ SFs in the thioredoxin fold class
- Phosphate binding (α/β)₈ barrels
- FAD-utilizing enzymes
- Small SFs: N-acetyl neuraminate lyase, oxalocrotonate tautomerase, glyoxalase II
Complications for superfamily analysis

- Relationships are often too distant to easily determine membership in a superfamily
- Accurate clustering of subgroups/families within a superfamily is hard
- Distance metrics based on sequence or structure do not track cleanly with divergence of function
  - Connectivity between subgroups and families can be uneven and difficult to evaluate
  - Different families/subgroups evolve at different rates
  - A given function may have evolved more than once and by different paths
Conclusions

> Systems for functional representation such as the E.C. system may not correlate well with structural similarities, especially for superfamilies with highly divergent members.

> It is valuable to know the evolutionary model for functional divergence associated with a given superfamily as that defines how conserved characteristics of function correlate with conserved sequence/structural elements.

> In functionally diverse enzyme superfamilies, function representation should be structured at the levels of granularity that match the level of conservation in sequence & structure:
  - members of **FAMILIES** perform the same overall reaction with the same [sub]set of substrates/products using the same mechanism;
  - members of **SUPERFAMILIES** conserve only an aspect of function, e.g., a common partial rxn, or chemical capability which is mediated by the subset of active site residues common to the entire group.

> Superfamilies representing homologous but highly diverse members can be difficult to cluster into subgroups and families based on distance metrics measuring sequence & structure similarity.
Thanks!

**Babbitt Lab**

Corey Adams  
Holly Atkinson  
Shoshana Brown PhD  
Ranyee Chiang  
Nima Fayazmanesh  
Margaret Glasner PhD  
Courtney Harper  
Elaine Meng PhD  
Ray Nagatani  
Sunil Ojha PhD  
Mark Peterson  
Scott Pegg PhD  
Benjamin Polacco  
Grant Shackleford  
Alexandra Schnoes  
Jennifer Seffernick PhD (UCSF & Univ. of MN)

**Collaborators**

**UCSF**

UCSF Resource for Biocomputing, Informatics & Visualization (RBVI)  
Tom Ferrin, Director  
Conrad Huang, Co-Investigator  
Matt Jacobson  
Andrej Sali  
Brian Shoichet

*John Gerlt (Univ. of Illinois)*

Shelley Copley (Univ. of Colorado)  
Larry Wackett (Univ. of MN)  
Frank Raushel (Texas A&M)  
Steve Almo (Einstein)  
Walter Novak (Post-doc now at Brandeis)

$$ NIH, NSF, California Institute for Quantitative Biomedical Research (QB3)$$